

# A FIXED-POINT TYPE FOR OCTAVE

*David Bateman, Laurent Mazet, Véronique Buzenac-Settineri and Markus Muck*

Motorola Labs Paris  
Parc Les Algorithmes, Saint Aubin  
91193 Gif-Sur-Yvette Cedex - France

David.Bateman@motorola.com, Laurent.Mazet@motorola.com,  
Veronique.Buzenac@motorola.com, Markus.Muck@motorola.com

## ABSTRACT

This paper announces the availability of a fixed point toolbox for the *Matlab* compatible software package *Octave*. This toolbox is released under the GNU Public License, and can be used to model the losses in algorithms implemented in hardware. Furthermore, this paper presents as an example of the use of this toolbox, the effects of a fixed point implementation on the precision of an *OFDM* modulator.

## 1. INTRODUCTION

When implementing algorithms in hardware, it is common practice to reduce the accuracy of the representation of numbers to a smaller number of bits. This allows much lower complexity in the hardware, at the cost of accuracy and potential overflow problems. Such representations are known as fixed point [1].

Many previous authors have presented solutions for modelling fixed point representations [2, 3, 4]. A common point of all of these solution is that they are written in a low-level programming language such as *C* or *C++*. In addition the only code that is publically available is that presented by Kim et al [2], and it is released under a license limiting its use to academic use only.

However, being low level implementations of the fixed point representations, they don't allow the engineer developing an algorithm the freedom to easily test multiple ideas and their consequence on their fixed point implementations. This gap between an algorithms development and its implementation can result in implementations that are overly complex and/or the choice of the algorithm to implement being sub-optimal.

For this reason it is important to consider the implementation losses in algorithms, due to fixed point representations, early in their design. To do this, there is a clear need for support of the analysis of fixed point types in standard high-level engineering design tools. One such tool

that is used by most engineers is *Matlab* [5] or its open-source cousin *Octave* [6]. *Matlab* supports fixed point types through its *Simulink* package [7].

However, the authors of this paper have preferred to write a toolbox implementation of a fixed point type to ease the softwares using for those not having *Matlab* licenses, but equally for issues of the softwares use on parallel computers. The authors have equally made this toolbox available under the GNU Public License as part of the *Octave-Forge* package [8]. To the authors knowledge this is the first time a fixed point toolkit for a high level engineering tool is available that intrinsically treats real, complex and matrix fixed point types. It is equally the first time that such code is publicly available that allows its use for the development of algorithms in commercial applications, with the sole restriction that changes to the fixed point type itself are returned to the community.

This article, describes the contents and use of the fixed point toolbox, giving useful examples and limitations. In addition, as an example of the use of this toolbox, we analyse the effects of a fixed-point implementation on an *OFDM* modulator.

## 2. DESCRIPTION OF THE CODE

### 2.1. Representation of Fixed Point Numbers

Fixed point numbers can be represented digitally in several manners, including *sign-magnitude*, *ones-complement* and *twos-complement*. However, the *twos-complement* representation simplifies the implementation of many operators in hardware, and so it is most common to see the *twos-complement* representation uses. This toolbox therefore represents fixed point objects using the *twos-complement* representation. All fixed point objects in this toolbox are represented by a *long int* that is used in the following manner

- 1 bit representing the sign,

- $is$  bits representing the integer part of the number, and
- $ds$  bits representing the decimal part of the number.

The numbers that can then be represented are then given by

$$-2^{is} \leq x \leq 2^{is} - 1 \quad (1)$$

and the distance between two values of  $x$  that are not represented by the same fixed point value is  $2^{-ds}$ .

The number of bits that can be used in the representation of the fixed point objects is determined by the number of bits in a *long int* on the platform. Valid values include 32- and 64-bits. However, to simplify their lives, the authors have also chosen to reduce the available number of bits by 1, so that issues with the differences between representations of *long int* and *unsigned long int* need not be taken into account. Therefore valid values of  $is$  and  $ds$  are given by

$$0 \leq (is + ds) \leq n - 2 \quad (2)$$

where  $n$  is either 32 or 64, depending on the number of bits in a *long int*. It should be noted that given the above criteria it is possible to create a fixed point representation that lacks a representation of the number 1. This makes the implementation of certain operators difficult, and so the valid representations are further limited to

$$0 \leq (is, ds, is + ds) \leq n - 2 \quad (3)$$

This does not mean that other numbers can not be represented by this toolbox, but rather that the numbers must be scaled prior to their being represented.

This toolbox allows both fixed point real and complex scalars to be represented, as well as fixed point real and complex matrices. The real and imaginary parts of the fixed point numbers and each element of fixed point matrices having their own fixed point representation.

## 2.2. Creation of Fixed Point Numbers

Before using a fixed point type, some variables must be created that use this type. This is done with the function *fixed*. The function *fixed* can be used in several manners, depending on the number and type of the arguments that are given. It can be used to create scalar, complex, matrix and complex matrix values of the fixed type.

The generic call to *fixed* is *fixed(is, ds, f)*, where the variables  $is$ ,  $ds$  are as previously described. The variable  $f$  can be either a scalar, complex, matrix or complex matrix of values that will be converted to a fixed point representation. It can equally be another fixed point value, in which case *fixed* has the effect of changing the representation of  $f$  to another representation given by  $is$  and  $ds$ .

If matrices are used for  $is$ ,  $ds$ , or  $f$ , then the dimensions of all of the matrices must match. However, it is valid to have  $is$  or  $ds$  as scalar values, which will be expanded to the same dimension as the other matrices, before use in the conversion to a fixed point value. The variable  $f$  however, must be a matrix if either  $is$  or  $ds$  is a matrix.

The most basic use of the function *fixed* can be seen in the example

```
octave:1> a = fixed(7, 2, 1)
ans = 1
octave:2> isfixed(a)
ans = 1
octave:3> typeinfo(a)
ans = fixed scalar
```

which demonstrates the creation of a real scalar fixed point value with 7 bits of precision in the integer part, 2 bits in the decimal part and the value 1. The function *isfixed* can be used to identify whether a variable is of the fixed point type or not. Equally, using the *typeinfo* or *whos* function allows the variable to be identified as "fixed scalar".

Other examples of valid uses of *fixed* are

```
octave:1> a = fixed(7, 2, 1);
octave:2> b = fixed(7, 2+1i, 1+1i);
octave:3> c = fixed(7, 2, 255*rand(10, 10) - 128);
octave:4> is = 3 * ones(10,10) + 4*eye(10);
octave:5> d = fixed(is, 1, eye(10));
octave:6> e = fixed(7, 2, 255*rand(10, 10) -
> 128 + 1i*(255 * rand(10, 10) - 128));
```

With two arguments given to *fixed*, it is assumed that  $f$  is zero or a matrix of zeros, and so *fixed* called with two arguments is equivalent to calling with three arguments with the third argument being zero. For example

```
octave:1> a = fixed([7, 7], [2, 2], zeros(1,2));
octave:2> b = fixed([7, 7], [2, 2]);
octave:3> assert(a == b);
```

Called with a single argument *fixed*, and a fixed point argument,  $b = \text{fixed}(a)$  is equivalent to  $b = a$ . If  $a$  is not itself fixed point, then the integer part of  $a$  is used to create a fixed point value, with the minimum number of bits needed to represent it. For example

```
octave:1> b = fixed(1:4);
```

creates a fixed point row vector with 4 values. Each of these values has the minimum number of bits needed to represent it. That is  $b(1)$  uses 1 bit to represent the integer part,  $b(2:3)$  use 2 bits and  $b(4)$  uses 3 bits. The single argument used with *fixed* can equally be a complex value, in which case the real and imaginary parts are treated separately to create a composite fixed point value.

### 2.3. Overflow Behavior of the Fixed Point Type

When converting a floating point number to a fixed point number the overflow behavior of the fixed point type is such that it implements clipping of the data to the maximum or minimum value that is representable in the fixed point type. This effectively simulates the behavior of an analog to digital conversion. For example

```
octave:1> a = fixed(7, 2, 200)
a = 127.75
```

However, the overflow behavior of the fixed point type is distinctly different if the overflow occurs within a fixed point operation itself. In this case the excess bits generated by the overflow are dropped. For example

```
octave:1> a = fixed(7, 2, 127) + fixed(7, 2, 2)
a = -127
octave:2> a = fixed(7, 2, -127) + fixed(7, 2, -2)
a = 127
```

The case where the representation of the fixed point object changes is different again. In this case the sign is maintained, while the most-significant bits of the representation are dropped. For example

```
octave:1> a = fixed(6, 2, fixed(7, 2, -127.25))
a = -63.25
octave:2> a = fixed(6, 2, fixed(7, 2, 127.25))
a = 63.25
octave:3> a = fixed(7, 1, fixed(7, 2, -127.25))
a = -127.5
octave:4> a = fixed(7, 1, fixed(7, 2, 127.25))
a = 127
```

In addition to the overflow issue discussed above, it is important to take into account what happens when an operator is used on two fixed point values with different representations. For example

```
octave:1> a = fixed(7, 2, 1);
octave:2> b = fixed(6, 3, 1);
octave:3> c = a + b;
octave:4> d = [c.int, c.dec]
d =
    7    3
```

as can be seen the representation of the output fixed point value is promoted such that  $c.int = \max(a.int, b.int)$  and  $c.dec = \max(a.dec, b.dec)$ . If this promotion causes the maximum number of bits in a fixed point representation to be exceeded, then an error will occur.

### 2.4. Analysis of Complexity

After the fixed point type is first used, four variables are initialized. The *fixed\_point\_count\_operations* variable is of particular interest. The *Octave* fixed point type can keep track of all of the fixed point operations and their type. This

is very useful for a simple complexity analysis of the algorithms. To allow the fixed point type to track operations the variable *fixed\_point\_count\_operations* must be non-zero. The function *reset\_fixed\_operations*, can be used to reset the number of operations since the last reset as given by the *display\_fixed\_operations* function.

### 2.5. Accessing Internal Fields

Once a variable has been defined as a fixed point object, the parameters of the field of this structure can be obtained by adding a suffix to the variable. Valid suffixes are '.x', '.sign', '.int' and '.dec', which return

- .x* The floating point representation of the fixed point number
- .sign* The sign of the fixed point number
- .int* The number of bits representing the integer part of the fixed point number
- .dec* The number of bits representing the decimal part of the fixed point number

As each fixed point value in a matrix can have a different number of bits in its representation, these suffixes return objects of the same size as the original fixed point object. For example

```
octave:1> a = [-3:3];
octave:2> b = fixed(7, 2, a);
octave:3> b.sign
ans =
    -1    -1    -1     0     1     1     1
octave:4> b.int
ans =
     7     7     7     7     7     7     7
octave:5> b.dec
ans =
     2     2     2     2     2     2     2
octave:5> c = b.x;
octave:6> typeinfo(c)
ans = matrix
```

The suffixes *.int* and *.dec* can also be used to change the internal representation of a fixed point value. This can result in a loss of precision in the representation of the fixed point value, which models the same process as occurs in hardware. For example

```
octave:1> b = fixed(7, 2, [3.25, 3.25]);
octave:2> b(1).dec = [0, 2];
b =
     3    3.25
```

However, the value itself should not be changed using the suffix *.x*.

## 2.6. Function Overloading

An important consideration in the use of the fixed point toolbox is that many of the internal functions of *Octave*, such as *diag*, can not accept fixed point objects as an input. This package therefore uses the *dispatch* function of *Octave-Forge* to overload the internal *Octave* functions with equivalent functions that work with fixed point objects, so that the standard function names can be used. However, at any time the fixed point specific version of the function can be used by explicitly calling its function name. There are too many functions available for use with the fixed point type to list in this article, and so interested readers are referred to the software package itself [8]

## 2.7. Putting it all Together

Now that the basic functioning of the fixed point type has been discussed, it is time to consider how to put all of it together. The main advantage of this fixed point type over an implementation of specific fixed point code, is the ability to define a function once and use as either fixed or floating point. Consider the example

```
function [b, bf] = testfixed(is, ds, n)
a = randn(n, n);
af = fixed(is, ds, a);
b = myfunc(a, a);
bf = myfunc(af, af);
endfunction

function y = myfunc(a, b)
y = a + b;
endfunction
```

In this case *b* and *bf* will be returned from the function *testfixed* as floating and fixed point types respectively, while the underlying function *myfunc* does not explicitly define that it uses a fixed point type. This is a major advantage, as it is critical to understand the loss of precision in an algorithm when converting from floating to fixed point types for an optimal hardware implementation. This mixing of functions that treat both floating and fixed point types can even apply to Oct-files (The *Octave* equivalent of a *Matlab* mex-file), as will be discussed later.

The main limitation to the above is the use of the concatenation operator, such as *[a,b]*, that is hard-coded into current versions of *Octave* and is thus not aware of the fixed-point type. Therefore, such code should be avoided and the function *concat* supplied with this package used instead.

## 3. THE FIXED POINT TYPE APPLIED TO AN OFDM MODULATOR

As an example of the use of the fixed point toolbox applied to a real signal processing example, we consider the

implementation of a Radix-4 *IFFT* in an *OFDM* modulator [9]. Code for this *IFFT* has been written as a C++ template class, and integrated as an *Octave* Oct-file. This allowed a single version of the code to be instantiated to perform both the fixed and floating point implementations of the same code. The code for this example is available as part of the release of this software package.

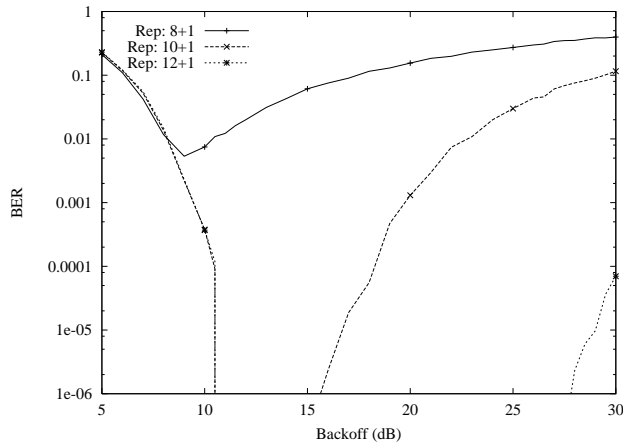
A particular problem of a hardware implementation of an *IFFT* is that each butterfly in the radix-4 *IFFT* consists of the summation of four terms with a suitable phase. Thus, an additional 2 output bits are potentially needed after each butterfly of the radix-4 *IFFT*. There are several ways of addressing this issue

- Increase the number of bits in the fixed point representation by two after each radix-4 butterfly. There are then two ways of treating these added bits
  - Accept them and let the size of the representation of the fixed point numbers grows. For large *IFFT*'s this is not acceptable
  - Cut the least significant bits of representation, either after each butterfly, or after groups of butterflies. This reduces the number of bits in the representation, but still trades off complexity to avoid an overflow condition
- Keep the fixed point representation used in the *IFFT* fixed, but reduce the input signal level to avoid overflows. The *IFFT* can then have internal overflows.

An overflow will cause a bit-error which is not necessarily critical. The last option is therefore attractive in that it allows the minimum complexity in the hardware implementation of the *IFFT*. However, careful investigation of the overflow effects are needed, which can be performed with the fixed point toolbox discussed in this article.

The figure 1 below shows the case of a 64QAM *OFDM* signal similar to that used in the 802.11a standard. In this figure only the *OFDM* modulator has been represented using fixed point, while the rest of the system is assumed to be perfect. Figure 1 shows the tradeoff between the backoff of the RMS power in the frequency domain signal relative to the fixed point representation for several different fixed point representations.

Two regions are clearly visible in figure 1. When the backoff of the RMS power is small, the effects of the overflow in the *IFFT* dominate, and reduce the performance. When the backoff is large, there are fewer bits in the fixed point representation relative to the average signal power and therefore a slow degradation in the performance. It is clear that somewhere between 11 and 13 bits in the representation of the fixed point numbers in the *IFFT* is optimal, with a backoff of approximately 13dB.



**Fig. 1.** Bit-error rate due to fixed point representation for various backoffs of the RMS power in frequency domain signal. Fixed point representation of  $N$  bits plus 1 bit for the sign

#### 4. CONCLUSION

This article has announced the release of a public available package for the analysis of fixed point implementations of algorithms within *Octave*. The code is available under the conditions of the GNU Public License. The basic capabilities of this code has been discussed and the simple examples of the code have been given.

Furthermore, this article has discussed the use of this package for the example of an *OFDM* modulator using a particular radix-4 *IFFT* implementation. The relationship between the clipping of the input signal, the number of bits in the fixed point representation and the noise introduced into the signal has been discussed.

#### 5. REFERENCES

- [1] W. Sung, "An automatic scaling method for the programming of fixed-point digital signal processors," in *Circuits and Systems, 1991 IEEE International Symposium on*, June 1991, pp. 11–14.
- [2] S. Kim, K. Kum, and W. Sung, "Fixed-point optimization utility for C and C++ based digital signal processing programs," in *Workshop on VLSI and signal processing*, 1995, pp. 197–206.
- [3] H. Keding, M. Willems, M. Coors, and H. Meyr, "FRIDGE: a fixed-point design and simulation environment," in *Design, Automation and Test in Europe, 1998., Proceedings*, Feb. 1998, pp. 429–435.
- [4] M. Harton and K. Kapuscinski, "BEC++: a software tool for increased flexibility in algorithm development," in *Speech Coding Proceedings, 1999 IEEE Workshop on*, June 1999, pp. 67–69.
- [5] Mathworks, "Matlab website," <http://www.mathworks.com>.
- [6] J. W. Eaton, "Octave website," <http://www.octave.org>.
- [7] Mathworks, "Fixed point blockset 4.1," <http://www.mathworks.com/products/fixpoint/>.
- [8] SourceForge, "Octave-Forge website," <http://octave.sourceforge.net>.
- [9] S. Gifford, J.E. Kleider, and S. Chuprun, "Broadband ofdm using 16-bit precision on a sdr platform," in *Military Communications Conference 2001*, 2001, vol. 1, pp. 180–184.